CEF

The Linear Regression Model

Estimation

The Ordinary Least Squares Estimator(OLS)

Goodness of Fit

The Standard Error of Regression

Prediction Using OLS

Assumptions

Variance of OLS Estimator

Confidence Interval

Homoskedastic v.s. Heteroskedastic

References

Regression with Single Variable [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

June 18, 2021

[1] This section is based on Stock and Watson (2020), Chapter 4 and 5, Hansen (2021), Chapter 2

CEF

The Linear Regression Model

Estimation

The Ordinary Least Squares Estimator(OLS)

Goodness of Fit

The Standard Error of Regression

Prediction Using OLS

Assumptions

Variance of OLS Estimator

Confidence Interval

Homoskedastic v.s. Heteroskedastic

References

Goal:

- Causal Inference
- Prediction

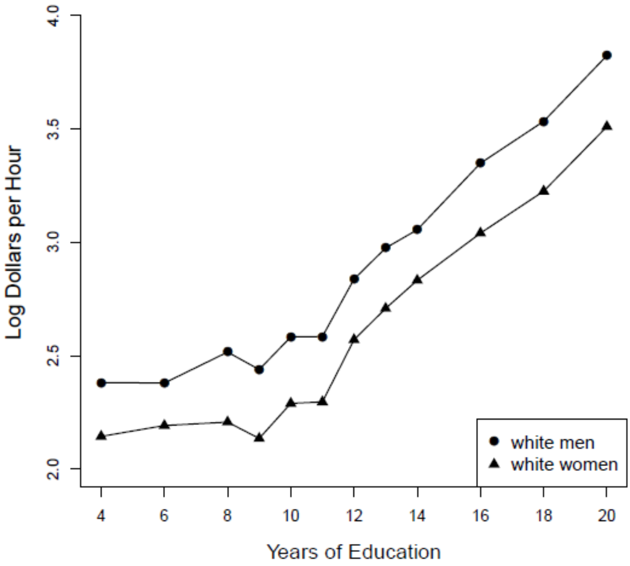
Data : Dependent variable Y, independent variable X.
Question of interest: How does change in X affect Y?

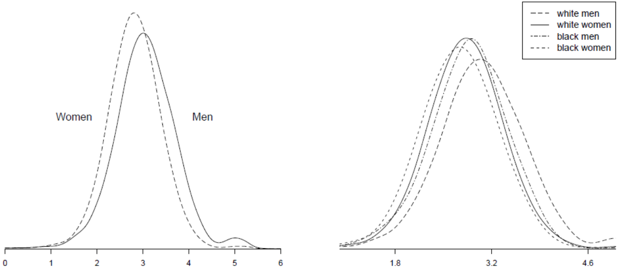
Example 1 I

An important determinant of wages is education[2]. In many empirical studies economists measure education attainment by number of years of schooling. Then the conditional expectation of $\log(wage)$ given *gender*, *race* and *education*, is a single number for each category.

$$\mathbb{E}(\log(wage)|gender = man, race = white, education = 12) = 2.8$$

[2] Population survey description `https://www.ssc.wisc.edu/~bhansen/econometrics/cps09mar_description.pdf` Data: `https://www.ssc.wisc.edu/~bhansen/econometrics/cps09mar.xlsx`


Log Dollars per Hour
2.0
2.5
3.0
3.5
4.0
4
6
8
10
12
14
16
18
20
Years of Education
white men
white women



(a) Women and Men

(b) By Gender and Race

CEF

The Conditional expectation can be written with

$$\mathbb{E}(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = m(x_1, \ldots, x_k).$$

We call this the **conditional expectation function(CEF)**.
The variables X can be both discrete and continuous.

Law of Iterated Expectation

If $\mathbb{E}(Y) < \infty$, then for any random variable X,

$$\mathbb{E}[\mathbb{E}[Y|X]] \ \mathbb{E}[Y].$$

More generally,
If $\mathbb{E}(Y) < \infty$, then for any random variables X_1 and X_2,

$$\mathbb{E}_{X_2}[\mathbb{E}[Y|X_1, X_2]] \ \mathbb{E}[Y|X_1].$$

CEF error I

Assume now X is the class size, and Y is expected test score for a given district.
The CEF error e is defined as the difference between Y and the *CEF* evaluated at X:

$$e = Y - m(X)$$

By construction, $Y = m(X) + e$.

CEF error II

A key property of CEF error is that it has conditional mean of zero.

$$\begin{aligned} \mathbb{E}(e|X) &= \mathbb{E}[(Y - m(X))|X] \\ &= \mathbb{E}[Y|X] - \mathbb{E}(m(X)|X) \\ &= m(X) - m(X) = 0. \end{aligned} \tag{1}$$

Properties of the CEF error, if $\mathbb{E}[Y] < \infty$ then

1. $\mathbb{E}[e|X] = 0$.
2. $\mathbb{E}[e] = 0$.
3. For any function $h(x)$ such that $\mathbb{E}[h(X)e] < \infty$ then $\mathbb{E}[h(X)e] = 0$.

CEF

The Linear Regression Model

Estimation

The Ordinary Least Squares Estimator(OLS)

Goodness of Fit

The Standard Error of Regression

Prediction Using OLS

Assumptions

Variance of OLS Estimator

Confidence Interval

Homoskedastic v.s. Heteroskedastic

References

CEF error III

A special case of the regression model is when there are no regressors X. In this case $m(X) = \mathbb{E}[Y] = \mu_Y$.
We can write the equation for Y in the regression format

$$Y = \mu + e, \mathbb{E}[e] = 0 \tag{2}$$

Variance I

An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error e.
We write this as

$$\sigma^2 = var[e] = \mathbb{E}[(e - \mathbb{E}(e))^2] = \mathbb{E}[e^2].$$

Consider the following regression:

$$Y = \mathbb{E}[Y|X] + e$$

- It turns out that there is a simple relationship. We can think of the conditional expectation $\mathbb{E}[Y|X]$ as the *"explained portion"* of Y.
- The remainder $e = Y - \mathbb{E}[Y|X]$ is the *"unexplained portion"*.

Variance II

Remark 1

In our discussion of iterated expectations we have seen that by increasing the conditioning set the conditional expectation reveals greater detail about the distribution of Y. What is the implication for the regression error?

More included variables indicate larger explained portion.

Best Predictor I

Suppose that given a random vector X we want to predict or forecast Y. We can write any predictor as a function g (X) of X. The (ex-post) prediction error is the realized difference $Y - g(X)$. A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}[(Y - g(X))^2].$$

We can define the best predictor as the function $g(X)$ which minimize the expectation of squares.

- The CEF $m(X)$ is the best predictor.
- If we assume no variation of X, then the best predictor is $\bar{Y}$.
- If we assume single X and linear function of $g(X)$, the best predictor is $\hat{\beta}_0 + \hat{\beta}_1 X$.

Best Predictor II

Theorem:
If $\mathbb{E}(Y^2) < \infty$, then for any predictor $g(X)$,

$$\mathbb{E}[(Y - g(X))^2] \geq \mathbb{E}[(Y - m(X))^2],$$

where $m(X) = \mathbb{E}[Y|X]$.
[3]

[3] For reference reading, see Hansen (2021), Chapter 2.

Conditional Variance I

While the conditional mean is a good measure of the location of a conditional distribution it does not provide information about the spread of the distribution.
A common measure of the dispersion is the conditional variance.
The conditional regression error given $X = x$ is

$$\sigma^2(x) = var[W|X = x] = \mathbb{E}[(e - \mathbb{E}[e|X = x])^2|X = x] = \mathbb{E}[e^2|X = x].$$

The conditional variance is a random variable.
we define the conditional standard deviation as its square root $\sigma(x) = \sqrt{\sigma^2(x)}$.

CEF

The Linear Regression Model

Estimation

The Ordinary Least Squares Estimator(OLS)

Goodness of Fit

The Standard Error of Regression

Prediction Using OLS

Assumptions

Variance of OLS Estimator

Confidence Interval

Homoskedastic v.s. Heteroskedastic

References

Conditional Variance II

The variance of Y can be decomposed as the following:

$$var(Y) = \mathbb{E}[var(Y|X) + var[\mathbb{E}[Y|X]]].$$

See Theorem 4.14 of Introduction to Econometrics. Theorem 2.8 decomposes the unconditional variance into what are sometimes called the "**within group variance**" and the "**across group variance**".

Issue with CEF

This assumption does not need to be viewed as literally true. Rather it is a useful modeling device so that parameters such as β are well defined.

- Some authors prefer the label the **data-generating-process (DGP)**. You can think of it as a theoretical concept or an infinitely-large potential population. In contrast, we refer to the observations available to us as the sample or dataset.
- Even in this case we view the observations as if they are random draws from an underlying infinitely-large population as this will allow us to apply the tools of statistical theory.

Linear CEF

An important special case is when the CEF $m(x) = \mathbb{E}[Y|X = x]$ is linear in x. In this case we can write the mean equation as

$$m(x) = \beta_0 + \beta_1 x_1.$$

Denote the vector $X = \begin{pmatrix} 1 \\ X_1 \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, then $m(x) = x^\top \beta$.

This is the **linear CEF** model. It is also often called the **linear regression model**, or the regression of Y on X.

Best Linear Predictor I

- While the conditional mean $m(X) = E[Y|X]$ is the best predictor of Y among all functions of X, its functional form is typically unknown.
- In particular, the linear CEF model is empirically unlikely to be accurate unless X is discrete and low-dimensional so all interactions are included.
- Consequently, in most cases it is more realistic to view the linear specification as an approximation.
- The conditional mean $m(X)$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.

For this derivation we require the following regularity condition.

1. $\mathbb{E}[Y^2] < \infty$
2. $\mathbb{E}\|X^2\| < \infty$
3. $\mathbf{Q}_{XX} = \mathbb{E}[XX^\top]$ is positive definite.

Example 2 I

For example [4]:

- A father tells you that his family wants to move to a town with a good school system. He is interested in a specific school district: Test scores for this district are not publicly available,
- the father knows its class size, based on the district's student–teacher ratio. So he asks you: if he tells you the district's class size, could you predict that district's standardized test scores?
- These two questions are clearly related: They both pertain to the relation between class size and test scores. Yet they are different. To answer the superintendent's question, you need an estimate of the causal effect of a change in one variable (the student-teacher ratio, X) on another (test scores, Y).

Example 2 II

- To answer the question, you need to know how X relates to Y, on average, across school districts so you can use this relation to predict Y given X in a specific district.
- Back to the question of school district, could you predict the district's standardized test scores?
- We use the notation $\mathbb{E}(Y|X=x)$ to denote the mean of Y given that X takes the value of x.
- In the case of test scores and class size, the linear function can be written $\mathbb{E}(TestScore|ClassSize) = \beta_0 + \beta_{ClassSize} \times ClassSize$, where β_0 is the intercept, and $\beta_{ClassSize}$ is the slope.
- Suppose the class size in the district size is 20, $\beta_0 = 720$ and $\beta_{ClassSize} = -0.6$. We could predict the mean test scores to be 720 -0.6 * 20 = 708.

Example 2 III

The prediction tells you what the test score will be, on average, for districts with class sizes of that value; it does not tell you what specifically the test score will be in any one district.

- Districts with the same class sizes can nevertheless differ in many ways and in general will have different values of test scores.
- If we make a prediction for a given district, we know that prediction will not be exactly right: The prediction will have an error.
- Stated mathematically, for any given district the imperfect relationship between class size and test score can be written as $TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + error$.
- $\beta_0 + \beta_{ClassSize} \times ClassSize$ represents the average relationship between class size and scores in the population of school districsts.
- *error* represents the error made in prediction.

[4] See Stock and Watson (2020) Chapter 4, example 1.

Notations

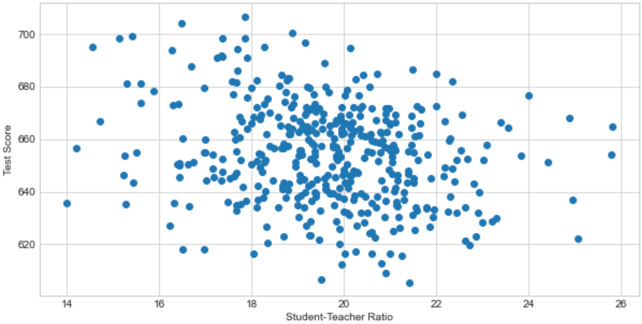
More generally, suppose we have n sample districts. Y_i denotes the average test score in i-th district and X_i be the average class size in i-th district. The prediction becomes $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_{1,i}$.

- the subscript i runs over observations $i = 1, \ldots, n$;
- Y_i is the dependent variable, the regressand or left-hand side variable(LHS).
- $X_{1,i}$ is the independent variable, the regressor or right-hand side variable(RHS).
- $\beta_0 + \beta_1 X_{1,i}$ is the population regression function.
- u_i is the error term.
- β_1 is the slope, β_0 is the intercept of the population regression function.

Textbook case

In a practical situation such as the application to class size and test scores, the intercept β_0 and the slope β_1 of the population regression line are unknown.
Therefore, we must use data to estimate these unknown coefficients.
This estimation problem is similar to the estimating sample mean.
`http://fmwww.bc.edu/ec-p/data/stockwatson/caschool.dta`

Scatter Plot



The Ordinary Least Squares Estimator I

The moment estimator of $\hat{S}(\beta)$ is the sample average:

$$\hat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 = \frac{1}{n}SSE(\beta)$$

where

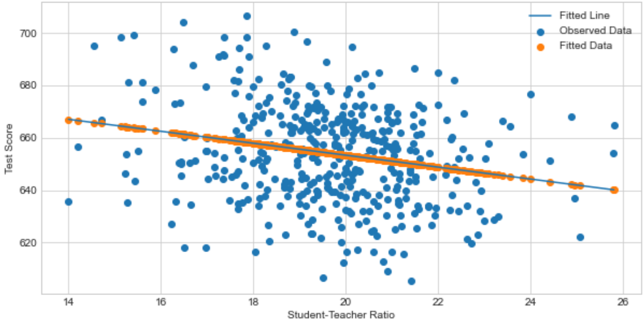
$$SSE(\beta) = \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2$$

is called the **sum of squared error** function.
The least squares estimator is $\hat{\beta} = \arg\min \hat{S}(\beta)$ where $\hat{S}(\beta)$ is defined above.

The Ordinary Least Squares Estimator II

The estimator is also commonly refered to as the **ordinary least squares (OLS)** estimator.

- It is important to understand the distinction between population parameters such as β and sample estimators such as $\hat{\beta}$.
- The population parameter β is a non-random feature of the population, is fixed,
- while the sample estimator $\hat{\beta}$ is a random feature of a random sample, varies across samples.


Fitted Line
Observed Data
Fitted Data
700
680
660
640
620
Test Score
14
16
18
20
22
24
26
Student-Teacher Ratio

Figure

Probability
Hao
CEF
The Linear Regression Model
Estimation
The Ordinary Least Squares Estimator(OLS)
Goodness of Fit
The Standard Error of Regression
Prediction Using OLS
Assumptions
Variance of OLS Estimator
Confidence Interval
Homoskedastic v.s. Heteroskedastic
References

Why OLS?

- OLS is the dominant method used in practice, it has become the common language for regression analysis throughout economics, finance
  - "The 'Beta' of a Stock" ,
  - and the social sciences more generally.
- Easy to use, build in most of the programming languages.

Solving for OLS with One Regressor

$$SSE(\beta) = \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 = \sum_{i=1}^{n} Y_i^2 - 2\beta\left(\sum_{i=1}^{n} X_i Y_i\right) + \beta^2\left(\sum_{i=1}^{n} X_i^2\right).$$

The OLS estimator $\hat{\beta}$ minimizes this function.

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

Note that the intercept-only model has $X_i = 1$. In this case $\hat{\beta} = \bar{Y}$.

Solving for OLS with Multiple Regressor I

To illustrate, consider a vector $X = (X_1, X_2)^\top$.

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 \\ &= \sum_{i=1}^{n} Y_i^2 - 2\beta^\top \left(\sum_{i=1}^{n} X_i Y_i \right) + \beta^\top \left(\sum_{i=1}^{n} X_i X_i^\top \right) \beta. \end{aligned} \tag{3}$$

A simple way to find the minimum is by solving the first order condition:

$$\frac{\partial}{\partial \beta} SSE(\hat{\beta}) = -2 \sum_{i=1}^{n} X_i Y_i + 2 \sum_{i=1}^{n} X_i X_i^\top \hat{\beta} = 0.$$

The solution for $\hat{\beta}$ may be found by solving the system of equation. We can write the solution using matrix algebra:

Solving for OLS with Multiple Regressor II

$$\sum_{i=1}^{n} X_i X_i^\top \hat{\beta} \sum_{i=1}^{n} X_i Y_i.$$

The system of equations of the form $\mathbf{Ab} = \mathbf{c}$ where A is $k \times k$ matrix and $\mathbf{b}$ and $\mathbf{c}$ are $k \times 1$ vectors is that $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$.
We can solve for the explicit formula for the least square estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right).$$

Best Linear Predictor I

The linear model applies to random variables (Y, X) can be viewed as a probability model. The model is

$$Y = X^\top \beta + e$$

where the linear coefficient β is defined as

$$\beta = \arg\min_{b \in \mathbb{R}^2} \mathbb{E}[(Y - X^\top b)^2].$$

The best linear predictor of Y given X for a pair of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^2$. We are interested in estimating the parameters β of the model, in particular the projection coefficient

$$\beta = (\mathbb{E}[XX^\top])^{-1}\mathbb{E}[XY].$$

Best Linear Predictor II

Notationally we wish to distinguish observations (realizations) from the underlying random variables.

- The random variables are (Y, X).
- The observations are (Y_i, X_i).
- From the vantage of the researcher the latter are numbers. From the vantage of statistical theory we view them as *realizations of random variables*.
  - For individual observations we append a subscript $i = 1, \ldots, n$
  - The number n is the sample size.
  - The dataset or sample is $\{(Yi, Xi) : i = 1, \ldots, n\}$. From the viewpoint of empirical analysis a dataset is an array of numbers.

R^2

The regression R^2 is the fraction of sample variance of Y explained by (or predicted by X).

$$\begin{aligned} ESS &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ TSS &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \end{aligned} \tag{4}$$

The ESS is the **explained sum of squares** and TSS is the **total sum of squares**.
$R^2 = \frac{ESS}{TSS}$.
The sum of squared residuals (SSR) is the sum of squared OLS residuals.
$R^2 = 1 - \frac{SSR}{TSS}$.

Probability

Hao

CEF

The Linear Regression Model

Estimation

The Ordinary Least Squares Estimator(OLS)

Goodness of Fit

The Standard Error of Regression

Prediction Using OLS

Assumptions

Variance of OLS Estimator

Confidence Interval

Homoskedastic v.s. Heteroskedastic

References

Variation of Data

- The standard error of regression(SER) is an estimator of the standard deviation of the regression error.
- $SER = \sqrt{s_{\hat{u}}^2}$ where $s_{\hat{u}}^2 = \frac{SSR}{n-2}$.
- The degree of freedom is $n-2$, because when two coefficients were estimated (β_0 and β_1).

Prediction

- **In-sample prediction** : The predicted value Y_i for the i-th observation is the value of Y_i predicted by the OLS regression line when X takes on its value X_i for that observation.
- **Out-of-sample prediction**: prediction methods are used to predict Y when X is known but Y is not.

Key Assumptions

Key Assumptions

1. The Error Term has Conditional Mean of Zero $\mathbb{E}[u_i|X_i] = 0$.
   The Error Term has Conditional Mean of Zero $\mathbb{E}[u_i|X_i] = 0$.
2. Independently and Identically Distributed Data (X_i, Y_i) are i.i.d.
3. Large Outliers are Unlikely $\mathbb{E}(X_i^4) < \infty, \mathbb{E}(Y_i^4) < \infty$.

Review: Distribution of $\bar{Y}$

Review of the sampling distribution of $\bar{Y}$.

- $\bar{Y}$ is an estimator of the unknown population mean of Y, μ_Y.
- $\bar{Y}$ is a random variable that takes on different values from one sample to the next; the probability of these different values is summarized in its sampling distribution.
- When sample size is small, the distribution follows a t-distribution with degree of freedom $n - 1$
- When sample size is large, the central limit indicate $\bar{Y}$ follows normal distribution.

Vectorized Terms

The n random observation can be viewed in vector term.

Let $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}$, $\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$.

Unbiasedness I

Claim: The OLS estimator is unbiased in the linear regression model. Consider the model where $Y = X_1\beta_1 + X_2\beta_2$.
This calculation can be done using either summation notation or matrix notation.

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}|X_1, X_2] &= \mathbb{E}[\left(\sum_{i=1}^{n} X_i X_i^{\top}\right) -1 \left(\sum_{i=1}^{n} X_i Y_i\right) |X_1, X_2] \\
&= \left(\sum_{i=1}^{n} X_i X_i^{\top}\right)^{-1} \mathbb{E}[\left(\sum_{i=1}^{n} X_i Y_i\right) |X_1, X_2] \\
&= \left(\sum_{i=1}^{n} X_i X_i^{\top}\right)^{-1} \sum_{i=1}^{n} \mathbb{E}[(X_i Y_i) |X_1, X_2] \\
&= \left(\sum_{i=1}^{n} X_i X_i^{\top}\right)^{-1} \sum_{i=1}^{n} X_i \mathbb{E}[(Y_i) |X_1, X_2] \\
&= \left(\sum_{i=1}^{n} X_i X_i^{\top}\right)^{-1} \sum_{i=1}^{n} X_i X_i^{\top} \beta \\
&= \beta.
\end{aligned} \tag{5}
$$

Unbiasedness II

If we write in matrix term, the expectation can be written as

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \begin{pmatrix} \vdots \\ \mathbb{E}[Y_i|\mathbf{X}] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}[X_i^\top \beta | X_i] \\ \vdots \end{pmatrix} = \mathbf{X}\beta. \tag{6}$$

Similarly

$$\mathbb{E}[\mathbf{e}|\mathbf{X}] == \begin{pmatrix} \vdots \\ \mathbb{E}[e_i|\mathbf{X}] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}[e_i|X_i] \\ \vdots \end{pmatrix} = 0. \tag{7}$$

Insert $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ into the formula for $\hat{\beta}$ to obtain

Unbiasedness III

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \left(\mathbf{X}\beta + \mathbf{e}\right)\right) \\ &= \beta + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{e}\right)\end{aligned} \tag{8}$$

Then

$$\mathbb{E}[\hat{\beta} - \beta | \mathbf{X}] = \mathbb{E}[\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{e}\right) | \mathbf{X}] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{e} | \mathbf{X}] = 0.$$

Conditional Variance I

For any $r \times 1$ random vector Z, we define the $r \times r$ covariance matrix

$$var(Z) = \mathbb{E}[((Z) - \mathbb{E}[Z])((Z) - \mathbb{E}[Z])^\top] = \mathbb{E}[ZZ^\top] - \mathbb{E}[Z]\mathbb{E}[Z]^\top.$$

and for any pair of (Z, X), the conditional covariance matrix

$$var[Z|X] = \mathbb{E}[((Z) - \mathbb{E}[Z|X])((Z) - \mathbb{E}[Z|X])^\top |X].$$

The variance of error vector $\mathbf{e}$ given $\mathbf{X}$ is a $n \times n$ matrix

$$var(\mathbf{e}|\mathbf{X}) = \mathbb{E}[\mathbf{e}\mathbf{e}^\top|\mathbf{X}] = \mathbf{D}.$$

The i-th diagonal element of $\mathbf{D}$ is

$$\mathbb{E}[e_i^2|\mathbf{X}] = \mathbb{E}[e_i^2|X_i] = \sigma_i^2,$$

while the i, j-th off diagonal element of $\mathbf{D}$ is

$$\mathbb{E}[e_i e_j|\mathbf{X}] = \mathbb{E}[e_i|X_i]\mathbb{E}[e_j|X_j] = 0,$$

Conditional Variance II

Under linear homoscedastic regression model, then $\mathbb{E}[e_i^2|X_i] = \sigma_i^2 = \sigma^2$.
For any $n \times r$ matrix $\mathbf{A} = \mathbf{A}(\mathbf{X})$,

$$var[\mathbf{A}(\mathbf{X})^\top \mathbf{Y}] = var[\mathbf{A}(\mathbf{X})^\top \mathbf{e}|\mathbf{X}] = \mathbf{A}^\top \mathbf{D}\mathbf{A}.$$

In particular, we write $\hat{\beta} = \mathbf{A}^\top Y$ where $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}$ and therefore

$$V_{\hat{\beta}} = \mathbf{A}^\top \mathbf{D}\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{D}\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}.$$

Note that under homoscedastic assumption, $\mathbf{X}^\top \mathbf{D}\mathbf{X} = \sigma^2 \mathbf{X}^\top \mathbf{X}$, then $V_{\hat{\beta}} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Intercept Only Model

Consider the intercept only model, where $Y = \beta_0 + \epsilon$.

- If $X_i = \begin{pmatrix} 1 \\ X_{i,1} \end{pmatrix}$, then $\mathbf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. We can compute the variance of $\hat{\beta}$ using the above formula.
- $\mathbf{X}^\top \mathbf{X} = n$, then $(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{n}$.
- Recall we estimate $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-1}$, the variance for $V_{\beta_0} = \frac{\hat{\sigma}^2}{n}$.

Single Regressor Model

Consider the one-regressor model $Y = \beta_0 + \beta_1 X_1 + \epsilon$.

- If $X_i = \begin{pmatrix} 1 \\ X_{i,1} \end{pmatrix}$, then $\mathbf{X} = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}$.
- We can compute the variance of $\hat{\beta}$ using the above formula.
- $\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$, then $(V_{\hat{\beta}} = \frac{\hat{\sigma}^2}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix}$.
- Recall we estimate $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}$,
- The diagonal terms corresponds to $V_{\hat{\beta}_0}$ and $V_{\hat{\beta}_1}$.

Hypothesis Testing

- Your client, the superintendent, calls you with a problem. She has an angry taxpayer in her office who asserts that cutting class size will not help boost test scores, so hiring more teachers is a waste of money. *ClassSize*, the taxpayer claims, has no effect on test scores.
- The taxpayer's claim can be restated in the language of regression analysis: $\beta_{ClassSize} = 0$.
- You already provided the superintendent with an estimate of $\beta_{ClassSize}$ using your sample of 420 observations on California school districts.
- Under the assumption that the least squares assumptions, Is there evidence in your data this slope is nonzero? Can you reject the taxpayer's hypothesis that $\beta_{ClassSize} = 0$?

Testing the slope I

We can construct a t-statistics from the data.

- Recall when testing the null hypothesis of the mean of Y equals to specific value: $H_0 : \mathbb{E}[Y] = \mu_Y$.
- The two-sided alternative hypothesis is $H_1 : \mathbb{E}[Y] \neq \mu_Y$.
- The t-test statistics is $t = \frac{\bar{Y} - \mu_{Y,0}}{s.e.(Y)}$.

We then compute the p-value by examine the distribution table.

Testing the slope II

At a theoretical level, the critical feature justifying the foregoing testing procedure for the population mean is that, in large samples, the sampling distribution of Y is approximately normal.

- Because $\hat{\beta}_1$ also has a normal sampling distribution in large samples, hypotheses about the true value of the slope β_1 can be tested using the same general approach.
- The null and alternative hypotheses need to be stated precisely before they can be tested. The hypothesis is that $\beta_{ClassSize} = 0$.
- More generally, under the null hypothesis the true population coefficient β_1 takes on some specific value $\beta_{1,0}$.
- Under the two-sided alternative, $H_1 : \beta \neq \beta_{1,0}$.

Steps to Test the Two-sided Hypothesis I

More generally, the **null hypothesis** and the **two-sided alternative hypothesis** are

$$H_0 : \beta_1 = \beta_{1,0} \ H_1 : \beta_1 \neq \beta_{1,0}.$$

The first step is to compute the **standard error** of $\hat{\beta}_1$. The slope estimator has the variance of

$$V_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2}.$$

The second step is to compute the $t-statistics$,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{V_{\hat{\beta}_1}}.$$

The third step is to compute the p-value, the probability of observing a value of $\hat{\beta}_1$ at least as different from $\beta_{1,0}$.

$$p-value = \mathbb{P}([|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{OLS} - \beta_1|]) = \mathbb{P}(|t| > |t^{OLS}|) = 2\Phi(-|t^{OLS}|$$

Steps to Test the Two-sided Hypothesis II

- A p-value of less than 5% provides evidence against the null hypothesis in the sense that, under the null hypothesis, the probability of obtaining a value of β_1 at least as far from the null as that actually observed is less than 5%. If so, the null hypothesis is rejected at the 5% significance level.
- Alternatively, the hypothesis can be tested at the 5% significance level simply by comparing the absolute value of the t-statistic to 1.96, the critical value for a two-sided test, and rejecting the null hypothesis at the 5% level if $|\hat{t}^{OLS}| > 1.96$.

Why Use Two-sided Test

- In practice, one-sided alternative hypotheses should be used only when there is a clear reason for doing so.
- This reason could come from economic theory, prior empirical evidence, or both. However, even if it initially seems that the relevant alternative is one-sided, upon reflection this might not necessarily be so.
- A newly formulated drug undergoing clinical trials actually could prove harmful because of previously unrecognized side effects.
- In the class size example, we are reminded of the graduation joke that a university's secret of success is to admit talented students and then make sure that the faculty stays out of their way and does as little damage as possible. In practice, such ambiguity often leads econometricians to use two-sided tests.

Confidence Interval for I

Because any statistical estimate of the slope β_1 necessarily has sampling uncertainty, we use the OLSestimator and its standard error to construct a confidence interval for the slope β_1 or for the intercept β_0.

- A 95 % two-sided confidence interval for β_1 is an interval that contains the true value of β_1 with a 95% probability;
- Equivalently, it is the set of values of β_1 that cannot be rejected by a 5 % two-sided hypothesis test.

When the sample size is large, it is constructed as

$$CI = [\hat{\beta} - Z_{1-\alpha/2} V_{\hat{\beta}_1}, \hat{\beta} + Z_{1-\alpha/2} V_{\hat{\beta}_1}]$$

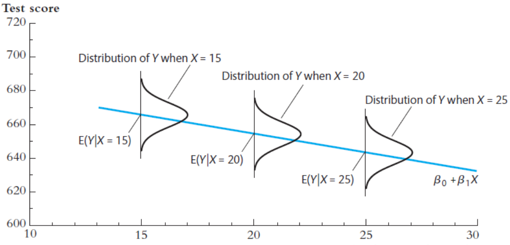
where $\alpha = 0.05$ and $Z_{1-\alpha/2} = 1.96$.

- The 95% confidence interval for β_1 can be used to construct a 95% confidence interval for the predicted effect of a general change in X.

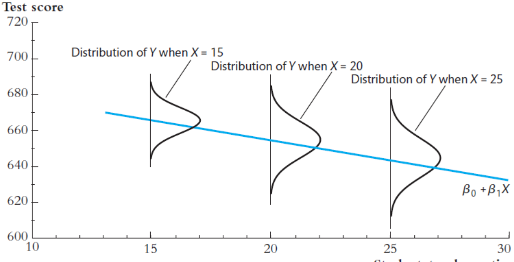
Homoskedastic v.s. Heteroskedastic I

The error term u_i is **homoskedastic** if the variance of the conditional distribution of e_i given X_i is constant for $i = 1, \ldots, n$ and in particular does not depend on X_i.
Otherwise, the error term is **heteroskedastic**.

Homoskedastic v.s. Heteroskedastic II



(a) The errors are homoscedastic



(b) The errors are heteroskedastic

Example

Higher be a binary variable that equals 1 for people whose father's NS-SEC grouping was higher than equals 0 if this grouping was routine.

$$Earnings_i = \beta_0 + \beta Higher_i + u_i$$

for $i = 1, \ldots, n$.

- The definition of homoskedasticity states that the variance of u_i does not depend on the regressor. Here the regressor is $Higher_i$, so at issue is whether the variance of the error term depends on $Higher_i$,
- In other words, is the variance of the error term the same for people whose father's socioeconomic classification was higher and for those whose father's socioeconomic classification was lower? If so, the error is **homoskedastic**; if not, it is **heteroskedastic**.

5

[5] This example is based on Stock and Watson (2020) p.p. 122.

Which error assumption to choose? I

1. Which is more realistic, heteroskedasticity or homoskedasticity? The answer to this question depends on the application.
   - ◇ Those who are born into relatively poorer circumstances are more likely to remain in poorer circumstances later in life, and live in households where earnings do not fall into the top income bracket.
   - ◇ In other words, the variance of the error term in for those whose father's socioeconomic classification was lower is plausibly less than the variance of the error term for those whose father's socioeconomic classification was higher.
   - ◇ Unless there are compelling reasons to the contrary—and we can think of none—it makes sense to treat the error term in this example as heteroskedastic.
   - ◇ It therefore is prudent to assume that the errors might be heteroskedastic unless you have compelling reasons to believe otherwise.

Which error assumption to choose? II

2. Practical implications.
   - In this regard, it is useful to imagine computing both, then choosing between them.
   - For simplicity, always to use the heteroskedasticity-robust standard errors.
   - Many software programs report homoskedasticityonly standard errors as their default setting, so it is up to the user to specify the option of heteroskedasticity-robust standard errors.

References I

Hansen, B. (2021). Introduction to econometrics. madison: University of wisconsin.

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.