Econometrics
Simple Linear Regression

Burcu Eke

UC3M

Linear equations with one variable

Recall what a linear equation is:

- $y = b_0 + b_1 x$ is a linear equation with one variable, or equivalently, a straight line.
- linear on x, we can think this as linear on its unknown parameter, i.e., $y = 1.3 + 3x$
- b_0 and b_1 are constants, b_0 is the y-intercept and b_1 is the slope of the line, y is the dependent variable, and x is the independent variable
- slope of a line being b_1 means for every 1 unit horizontal increase there is a b_1 unit vertical increase/decrease depending on the sign of b_1.

Linear equations with one variable

- A linear equation is of **deterministic** nature ⇒ outcomes are precisely determined without any random variation
- A given input will always produce the same output ⇒ perfect relationship
- In real life data, it is almost impossible to have such a prefect relationship between two variables. We almost always rely on rough predictions. One of our tools to do so is regression.

Empirical and Theoretical Relationships

- Economists are interested in the relationship between two or more economic variables $\Rightarrow$ at least bivariate populations
- The economic theory in general suggests relationships in functional forms (recall economic models). These relations are deterministic, such as $Y = f(X)$, or $Y = f\left(X_1, X_2, ...X_k\right)$
- A given input will always produce the same output $\Rightarrow$ perfect relationship

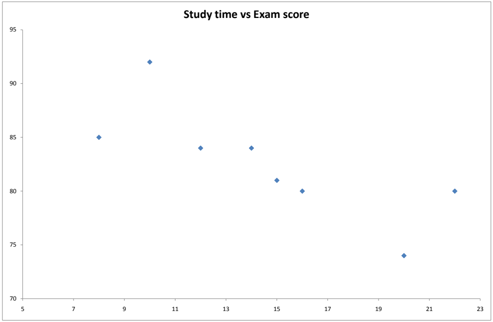
Example 1: Study time and score

Let's consider the following example. The table below shows the data for total hours studied for a calculus test, x, and the test scores, y.

Time (x)	Score (y)
10	92
15	81
12	84
20	74
8	85
16	80
14	84
22	80

Example 1, cont'd

Consider the scatter plot based on the data of example 1. Does it look like a perfect relationship?



Example 2: Savings and income

- Relationship between savings (Y) and income (X) (Goldberger, Chapter 1 of "A Course in Econometrics", Harvard U. Press. 1991)
  - Data from 1027 families, between 1960-1962, in the USA
- The joint distribution of savings and income are presented in the next table $\Rightarrow\ P(Y, X)$
- For discrete case, $P(Y, X) \ \Rightarrow\ P(Y = y \,\text{and}\, X = x)$.

Example 2, cont'd

Table : Joint Distribution Y, savings, and X, income: $P(Y, X)$

Y	X (in 1000 of Dollars) **1.4**	**3.0**	**4.9**	**7.8**	**14.2**	
(Savings rate)						(Sum
0.45	0.015	0.026	0.027	0.034	0.033	
0.18	0.019	0.032	0.057	0.135	0.063	
0.05	0.059	0.066	0.071	0.086	0.049	
-0.11	0.023	0.035	0.045	0.047	0.015	
-0.25	0.018	0.016	0.016	0.008	0.005	
P (X) (Sum of the Columns)	0.134	0.175	0.216	0.310	0.165	

Example 2, cont'd

- Among the information we can obtain from a joint probability table, there are two that are of interest to us:
  - whether we can have a deterministic relationship, i.e., $Y = f(X)$
  - whether savings and income are independent
- Can we have a deterministic relationship between savings and income based on the previous table?
  - **No**. In order to have a deterministic relationship, we need to have a unique savings for each level of income, in other words we need to have a functional relationship.
  - In terms of probability, for each income level, we need to have only one savings level with positive probability!

Example 2, cont'd

- But we do have a relationship between income level and savings $\Rightarrow$ As income increases, savings level increases.
- To further investigate this, lets calculate the conditional distribution:
  - $P(Y|X) = \dfrac{P(Y,X)}{P(X)}$
  - $\hat{\mu}_{Y|X=x} = \sum_{y} yP(Y = y|X = x)$

Example 2, cont'd

Table : Conditional Distribution of Y, savings, given X, income: $P(Y|X)$

	X (in 1000)				
Y	**1.4**	**3.0**	**4.9**	**7.8**	**14.2**
0.45	0.112	0.149	0.125	0.110	0.200
0.18	0.142	0.183	0.264	0.435	0.382
0.05	0.440	0.377	0.329	0.277	0.297
-0.11	0.172	0.200	0.208	0.152	0.091
-0.25	0.134	0.091	0.074	0.026	0.030
Column sum	1	1	1	1	1
$\hat{\mu}_{Y\|X}$	0.045	0.074	0.079	0.119	0.156

What is the relationship between the conditional mean and income level?

Empirical and Theoretical Relationships

- The empirical relationships between economic variables $\Rightarrow$ **not** deterministic, but **stochastic**
- To combine theory and data, one must interpret the economic theory in a different way
  - When the economic theory postulates that Y is a function of X, $Y = f(X)$, it implies that the expected value of Y is a function of X, $E[Y] = f(X)$
  - $Y = f(X) \Rightarrow$ deterministic, $E[Y] = f(X) \Rightarrow$ stochastic

Prediction

- When we are in a stochastic setting, we are in general interested in prediction, but how do we form our prediction?
- One way would be just to guess a number, but do you think it will be good? How can we assess that it is good?
- We are interested in finding the best way of **predicting**

Prediction

- Suppose we want to predict Y, using the information on X. Let $c(X)$ be a prediction
- Then, $U = Y - c(X)$ will be our prediction error. We want the expected value of the square of our prediction, $E[U^2]$ error to be as small as possible
  - Why $E[U^2]$?
  - Choose a prediction so that $E[U^2] = \sum_{y \in Y} (y - c(X))^2 \, P(y)$ is minimized

Prediction: The best constant prediction

- We do not always have bivariate, or multivariate case, sometimes we only have information on Y
- In these cases, our prediction does not depend on other variables, so our prediction can be denoted as c
- Then we choose a prediction such that we minimize $E[U^2] = \sum_{y \in Y} (y - c)^2 P(y) \Rightarrow c = E[Y] = \mu_Y$
- Average of Y is the **best constant predictor**

Exercise 2, cont'd

Y (Saving rate)	$P(Y)$
0.45	0.135
0.18	0.306
0.05	0.331
-0.11	0.165
-0.25	0.063

Then the best constant predictor for the saving rate is:

$$\begin{aligned} E[Y] &= 0.45 * 0.135 + 0.18 * 0.306 + 0.05 * 0.331 \\ &- 0.11 * 0.165 - 0.25 * 0.063 \\ &= 0.09848 = 9.85\% \end{aligned}$$

Prediction: The best linear prediction

- In the case that we have bivariate data (Y, X), say savings rate **and** income levels, then we can make use of the relationship between them to predict Y
- Linear prediction implies that our $c(X)$ is linear, i.e., $c(X) = c_0 + c_1 X$
- So, we need to find c_0 and c_1 such that $E[U^2] = \sum_{y \in Y} (y - c(X))^2 P(y)$ is minimized

Prediction: The best linear prediction

- By replacing $c(X)$ by $c_0 + c_1 X$, and solving for the minimization problem we get
  - $c_0 = \alpha_0 = E(Y) - \alpha_1 E(X) = \mu_Y - \alpha_1 \mu_X$
  - $c_1 = \alpha_1 = \dfrac{C(X,Y)}{V(X)} = \dfrac{\sigma_{XY}}{\sigma_X^2}$
- The function $\alpha_0 + \alpha_1 X$ is the **linear projection** (or the **best linear prediction**) of Y given $X \Rightarrow L(Y|X) = \alpha_0 + \alpha_1 X$

Example 2, cont'd

- To get the best linear prediction for savings rate, we need to calculate: $E[X], E[Y], E[XY], C(X,Y), E[X^2], V(X)$
- For $E[XY]$:

XY	P(XY)	XY	P(XY)	XY	P(XY)	XY	P(XY)
-3.55	0.005	-0.75	0.016	0.07	0.059	0.54	0.032
-1.95	0.008	-0.54	0.045	0.15	0.066	0.63	0.015
-1.56	0.015	-0.35	0.018	0.25	0.071	0.71	0.049
-1.23	0.016	-0.33	0.035	0.25	0.019	0.88	0.057
-0.86	0.047	-0.15	0.023	0.39	0.086	1.35	0.026

- $E[XY] = \sum_{i=1}^{5} \sum_{j=1}^{5} X_i Y_j P\left(XY = X_i Y_j\right) = 0.783$

Example 2, cont'd

- $E[X] = 1.4 * 0.134 + 3.0 * 0.175 + 4.9 * 0.216 + 7.8 * 0.310 + 14.2 * 0.165 = 6.532$
- $C(X, Y) = E[XY] - E[X]E[Y] = 0.782607 - 6.532 * 0.09848 = 0.13934$
- $E[X^2] = 1.42^2 * 0.134 + 3.02^2 * 0.175 + 4.92^2 * 0.216 + 7.82^2 * 0.310 + 14.22^2 * 0.165 = 59.155$
- $V(X) = E[X^2] - (E[X])^2 = 59.155 - 6.5322^2 = 16.488$
- $c_0 = \alpha_0 = E(Y) - \alpha_1 E(X) = 0.0433$
- $c_1 = \alpha_1 = \dfrac{C(X, Y)}{V(X)} = 0.00845$

Example 2, cont'd

- $L(Y|X) = 0.043278 + 0.008451X$
- The for discrete values of X, we get

$$L(Y|X) = \begin{cases} 0.043278 + 0.008451 * 3.0 = 0.069 & \text{if } X = 3.0 \\ 0.043278 + 0.008451 * 4.9 = 0.085 & \text{if } X = 4.9 \\ 0.043278 + 0.008451 * 7.8 = 0.1092 & \text{if } X = 7.8 \\ 0.043278 + 0.008451 * 14.2 = 0.1633 & \text{if } X = 14.2 \end{cases}$$

Best Prediction

- So far we considered best constant predictor, and best linear predictor
- Let's relax the linearity restriction on $c(X)$, i.e., $c(X)$ can be any function that minimizes $E[U^2]$
- The **best predictor** of Y then becomes the conditional expected value of Y, $E[Y|X]$
  - If the $E[Y|X]$ is linear, then $E[Y|X]$ and $L(Y|X)$ are the same. The reverse is NOT true!!!
  - If the $E[Y|X]$ is NOT linear, then $L(Y|X)$ is the **best linear approximation** to $E[Y|X]$

Example 2, cont'd

X (income) (in \$1000s)	c BCP	$L(Y\|X)$ BLP	$E[Y\|X]$ BP
1.4	0.0985	0.055	0.045
3.0	0.0985	0.069	0.074
4.9	0.0985	0.085	0.079
7.8	0.0985	0.1092	0.119
14.2	0.0985	0.1633	0.156

Prediction

- Whenever $L(Y|X) \neq E[Y|X]$, $L(Y|X)$ provides a good approximation to $E[Y|X]$, hence can be used in some circumstances
- However, $E[Y|X]$ characterizes conditional mean of Y given X, the $L(Y|X)$ does not $\Rightarrow$ $E[Y|X]$ can have causal interpretation, the $L(Y|X)$ can not

SIMPLE LINEAR REGRESSION (SLR)

Simple Linear Regression

- Our big goal to analyze and study the relationship between two variables
- One approach to achieve this is simple linear regression, i.e, $Y = \beta_0 + \beta_1 X + \varepsilon$
- While answering our question, a simple linear regression model addresses some issues:
  1. How to deal with the factors other than X that effects Y
  2. How to formulate the relationship between X and Y
  3. Whether our model captures a ceteris paribus relationship between X and Y

Simple Linear Regression

SLR Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

- $Y \Rightarrow$ Dependent variable, endogenous variable, response variable, regressand ...
- $X \Rightarrow$ Independent variable , exogenous variable, control variable, regressor ...
- $\boldsymbol{\beta} = (\beta_0, \beta_1) \Rightarrow$ Parameter vector, population parameters
- $\varepsilon \Rightarrow$ Error term

Simple Linear Regression: Assumptions

A1: Linear in parameters

- It implies that a unit change in X has the same effect on Y, independently of the initial value of X.
- SLR is linear in parameters:
  - The following are linear in parameters: $Y = \beta_0 + \beta_1 X + \varepsilon$, $Y = \beta_0 X^2 + \beta_1 \frac{Z}{logZ} + \varepsilon$
  - The following is NOT linear in parameters: $Y = \beta_1 X^{\beta_2} \varepsilon$
- Violations of this assumption are form of a “specification errors, such as nonlinearity (when the relationship is not linear in parameters)
- We can overcome the limitations this assumption imposes

Simple Linear Regression: Assumptions

A2: $\mathbf{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$

- For the all values of X, the mean error rate is same and equal to 0.
- $E[\varepsilon|X] = 0$ implies for any function $h()$, $C(h(X), \varepsilon) = 0$. Hence, ε is not correlated with any function of X
- In particular, $C(X, \varepsilon) = 0$

$$\begin{aligned} C(X, \varepsilon) &= E[X\varepsilon] - E[X]E[\varepsilon] \\ E[X\varepsilon] &= E[E[X\varepsilon|X]] = E[XE[\varepsilon|X]] = 0 \\ E(X)E(\varepsilon) &= 0 \end{aligned}$$

Simple Linear Regression: Assumptions

- $E[Y|X] = \beta_0 + \beta_1 X$, the conditional expectation function is linear, hence
  - $C(Y, X) = C\left(E[Y|X], X\right) = \beta_1 V(X) \Rightarrow \beta_1 = \dfrac{C(Y, X)}{V(X)}$
  - $E[Y] = E\left[E[Y|X]\right] = \beta_0 + \beta_1 E[X] \Rightarrow \beta_0 = E[Y] - \beta_1 E[X]$
- $E[Y|X] = L(Y|X) = \beta + 0 + \beta_1 X$, where $L(.)$ is the linear projection

Simple Linear Regression: Assumptions

A3: Conditional Homoscedasticity $\Rightarrow\; V(\varepsilon|X) = \sigma^2 \,\forall\, X$

- This implies $V(\varepsilon) = \sigma^2$
  - $V(\varepsilon|X) = E[\varepsilon^2|X] - E\left[(E[\varepsilon|X])\right] = E[\varepsilon^2|X] = \sigma^2$
  - $E[\varepsilon^2] = E\left[E[\varepsilon^2|X]\right] = \sigma^2$
  - $V(\varepsilon) = E[\varepsilon^2] - \left(E[\varepsilon]\right)^2 = \sigma^2$
- This also implies $V(Y|X) = \sigma^2$

Simple Linear Regression: Interpretation of Coefficients

- From **A1**, we have $Y = \beta_0 + \beta_1 X + \varepsilon$, if we take the conditional expectation of both sides we get

$$\begin{aligned} E[Y|X] &= E\left[\beta_0 + \beta_1 X + \varepsilon\right] \\ &= \beta_0 + \beta_1 E[X|X] + E[\varepsilon|X] (\text{by } \mathbf{A2}) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

- This implies: $\beta_1 = \dfrac{\Delta E[Y|X]}{\Delta X}$. That is, When X increases by one unit, Y, on average, increases by β_1 units.

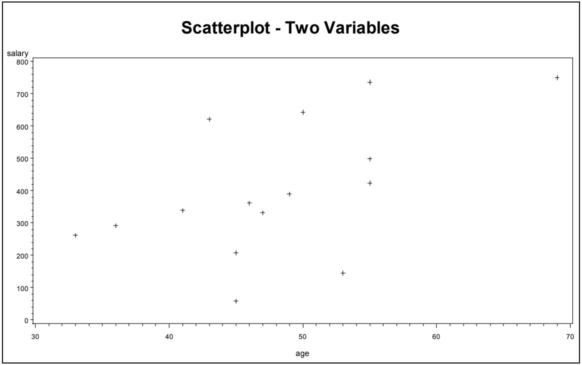
Simple Linear Regression: Interpretation of Coefficients

- $E[Y|X=0] = E[\beta_0 + \beta_1 X + \varepsilon | X = 0] = \beta_0$
- β_0 is the mean value of Y, when $X = 0$
- β_0 is the intercept term
- In the case that $X = 0$ is not included in the range of X in our data set, β_0 has no interpretation

Simple Linear Regression: Interpretation of Coefficients

- Notice that **A2** also implies $E[Y|X] = \beta_0 + \beta_1 X = L(Y|X)$, that is, the **best prediction** is the **best linear prediction**
- If **A2** is not satisfied, then $E[Y|X] \neq L(Y|X)$. This implies:
  - $\Rightarrow$ β_0 and β_1 are only the parameters of a linear projection, NOT of a conditional expectation
  - $\Rightarrow$ β_0 and β_1 do not have causal interpretations

Simple Linear Regression: Example 3



Simple Linear Regression: Example 3

- The regression equation: $Y = -226.53 + 13.10X$ (Assume that all assumptions are met)
- What is the interpretation of 13.10?
- Can we predict mean salary of a 0 years old person? Why? Why not?
- What is the interpretation of -226.53?

SIMPLE LINEAR REGRESSION: THE ESTIMATION

Simple Linear Regression: Estimation

- Our Objective: Estimate the population parameters, i.e., $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and σ^2
- Our Model: $Y = \beta_0 + \beta_1 X + \varepsilon$, where $E[\varepsilon|X] = 0$ and $V(\varepsilon|X) = \sigma^2$
- Our Data: Pairs of (y_i, x_i) $i = 1, 2, ..., n$ are the **sample** realizations of population data (Y_i, X_i):

y	x
y_1	x_1
y_2	x_2
.	.
.	.
.	.
y_n	x_n

Simple Linear Regression: Estimation

- We need a method to estimate our population parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and σ^2, using our model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for all $i = 1, 2, ..., n$

- We need to distinguish between population parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and σ^2 and the estimated values, denoted as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ and $\hat{\sigma}^2$

Estimation: The Analogy Principle

- An approach for estimation
- Use the corresponding sample quantity as an estimate of the population quantity
  - Use sample mean $\bar{Y} = \sum_{i=1}^{n} \frac{y_i}{n}$ to estimate the population mean
  - Use sample variance $s^2 = \sum_{i=1}^{n} \frac{\left(y_i - \bar{Y}\right)^2}{n-1}$ to estimate the population variance
- Method of moments: Let $\mu_i = E[X^i]$ be the i^{th} moment, the by method of moments, we estimate μ_i by corresponding sample moment $m_i = \sum_{k=1}^{n} \frac{(x_k)^i}{n}$

Estimation: The Analogy Principle

- Recall the bets prediction and the best linear prediction. Under our assumptions:
  - $E[Y|X] = L(Y|X) = \beta_0 + \beta_1 X$
  - $\beta_0 = E(Y) - \beta_1 E(X)$
  - $\beta_1 = \dfrac{C(X,Y)}{V(X)}$

Estimation: The Analogy Principle

- then, by using the analogy principle we have:
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\beta}_1 = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i^n (X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_i^n (X_i - \bar{X})Y_i}{\frac{1}{n}\sum_i^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$

Estimation: The Ordinary Least Squares (OLS)

- The idea of the (OLS) principle is to choose parameter estimates that minimize the squared error term, i.e., the distance between the data and the model
- $\min E[\varepsilon^2]$, or equivalently, $\min E\left[(Y - \beta_0 + \beta_1 X)^2\right]$
- $\varepsilon_i = Y_i - E[Y_i|X_i]$

The Ordinary Least Squares (OLS)

- For the sample data, we define the error term as

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- Then the sample equivalent of $\min E[\varepsilon^2]$ is

$$\min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

The Ordinary Least Squares (OLS)

- FOC are:
  - $\sum_{i=1}^{n} \hat{\varepsilon} = 0$
  - $\sum_{i=1}^{n} \hat{\varepsilon} X_i = 0$
- These FOC implies:
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n} c_i Y_i$ where $x_i = (X_i - \bar{X})$ and $c_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$

The Ordinary Least Squares (OLS)

- Compare Least square estimators and analogy principle estimators. are the similar?
- **Predicted values:** $\hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 X_i$
- Predicted values are orthogonal to the predicted error term $\Rightarrow \sum_i \hat{Y}_i \hat{\varepsilon}_i = 0$

Estimation Terminology

- **Estimator** is a RULE for calculating an estimate of a parameter based on observed data, i.e., $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- **Estimate** is a number that is an outcome of an estimation process, i.e, $\hat{\beta}_1 = 0.003$

The Properties of The OLS Estimators

P1 Linear in Y

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\beta}_1 = \sum_i c_i Y_i$

The Properties of The OLS Estimators

P2 Unbiasedness: Expected value of the estimator is equal to the parameter itself

- We achieve this property by **A1** and **A2**
- $E[\hat{\beta}_0] = \beta_0$
- $E[\hat{\beta}_1] = \beta_1$

The Properties of The OLS Estimators

The **unbiasedness** property indicates that if we have infinitely many samples of size-n from the same population and estimate the same model with each of the samples:

- We will have a distribution of the **estimator** of β_l, with a different realization for each sample
- The mean of that distribution will be the the population parameter β_l

The Properties of The OLS Estimators

$E[\hat{\beta}_1] = \beta_1$

$$
\begin{aligned}
\hat{\beta}_1 &= \sum_i c_i Y_i \\
&= \sum_i c_i \left(\beta_0 + \beta_1 X_i + \varepsilon_i\right) \\
&= \beta_0 \sum_i c_i + \beta_1 \sum_i c_i X_i + \sum_i c_i \varepsilon_i \\
&= \beta_1 \sum_i c_i X_i + \sum_i c_i \varepsilon_i \quad \left(\text{because } \sum_i c_i = 0\right) \\
&= \beta_1 + \sum_i c_i \varepsilon_i \quad \left(\text{because } \sum_i c_i X_i = 1\right)
\end{aligned}
$$

The Properties of The OLS Estimators

$$
\begin{aligned}
E[\hat{\beta}_1|X] &= E\left[\beta_1 + \sum_i c_i \varepsilon_i | X\right] \\
&= \beta_1 + \sum_i c_i \underbrace{E[\varepsilon_i|X]}_{0} \\
&= \beta_1 \\
E[\beta_1] &= E\left[E[\beta_1|X]\right] \\
&= E[\beta_1] = \beta_1
\end{aligned}
$$

The Properties of The OLS Estimators

The Variance

From **A1**, **A2**, and **A3**, we have

- $V(\hat{\beta}_0) = \sum_i \frac{X_i^2}{n} V(\hat{\beta}_1)$
- $V(\hat{\beta}_1) = \frac{\sigma^2}{n} E\left[\frac{1}{S_X^2}\right]$

The Properties of The OLS Estimators

Let's show that $V(\hat{\beta}_1) = \frac{\sigma^2}{n} E\left[\frac{1}{S_X^2}\right]$

$$
\begin{aligned}
V(\hat{\beta}_1) &= E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] = E\left[\left(\sum_i c_i \varepsilon_i\right)^2\right] = \sum_i E[c_i^2 \varepsilon_i^2] \\
&= E\left[\sum_i E\left[c_i^2 \varepsilon_i^2 | X\right]\right] = E\left[\sum_i c_i^2 E\left[\varepsilon_i^2 | X\right]\right] \\
&= \sigma^2 E\left[\sum_i c_i^2\right] \quad (\text{by } \mathbf{A3}) \\
&= \sigma^2 E\left[\sum_i \left(\frac{x_i}{\sum_i x_i^2}\right)^2\right] = \sigma^2 E\left[\sum_i \frac{\frac{1}{n}}{\frac{1}{n}(\sum_i x_i^2)^2}\right] \\
&= \frac{\sigma^2}{n} E\left[\frac{1}{S_X^2}\right]
\end{aligned}
$$

The Properties of The OLS Estimators: Gauss-Markov Theorem

- Under assumptions **A1** to **A3**, the Gauss-Markov Theorem states that the $\hat{\beta}_0$ and $\hat{\beta}_0$ are the efficient estimators among all linear unbiased estimators

- $\hat{\beta}_0$ and $\hat{\beta}_0$ are the most efficient estimators among all the linear unbiased estimators $\rightarrow$ Best Linear Unbiased Estimator

The Properties of The OLS Estimators: Gauss-Markov Theorem

- What does it mean to be a **consistent** estimator?
  - $\hat{\beta}_0$ and $\hat{\beta}_0$ are **consistent** if

$$p \lim_{n \to \infty} \hat{\beta}_j = \beta_j,\ j = 0, 1$$

    in other words,

$$\lim_{n \to \infty} P\left(\left|\hat{\beta}_j - \beta_j\right| < \delta\right) = 1\ \forall \delta > 0$$

- Consistency means the distributions of the estimators $\hat{\beta}_0$ and $\hat{\beta}_0$ become more and more concentrated near the true values β_0 and β_0 as sample size increases, so that the probability of $\hat{\beta}_0$ and $\hat{\beta}_0$ being arbitrarily close to β_0 and β_0 converges to one

- For these results, we need **A1** to **A3** to met!!!

The Variance of The OLS Estimators

- The variance of the OLS estimators,$\hat{\beta}_0$ and $\hat{\beta}_0$, depends on $\sigma^2 = V(\varepsilon) = E[\varepsilon^2]$

- The problem is, the errors $\varepsilon = \varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are not observed, so we need to estimate ε

$$\begin{aligned}\hat{\varepsilon} &= Y - \hat{\beta}_0 - \hat{\beta}_1 X \\ &= \beta_0 + \beta_1 X + \varepsilon - \hat{\beta}_0 - \hat{\beta}_1 X \\ &= \varepsilon - \left(\hat{\beta}_0 - \beta_0\right) - \left(\hat{\beta}_1 - \beta_1\right) X\end{aligned}$$

- Even though $E[\hat{\beta}_j] = \beta_j$ for $j = 0, 1$, $E[\hat{\varepsilon}] \neq \varepsilon$

The Variance of The OLS Estimators

- One attempt to estimate σ^2 would be to use $(1\backslash n)\sum\varepsilon^2$ instead of $E[\varepsilon^2]$. However, this is not feasible

- One attempt to estimate σ^2 would be using sample values of the error, $\hat{\varepsilon}$, and use this to estimate the sample variance: $\tilde{\sigma}^2 = (1\backslash n)\sum\varepsilon^2$. However, this is biased. The bias comes from unmatched degrees of freedom and denominator

- When we account for this bias, we get a **consistent** and **unbiased** estimate:

$$\hat{\sigma}^2 = \sum_i \frac{\varepsilon_i^2}{n-2}$$

The Variance of The OLS Estimators

- Note that both of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ are consistent and for larger samples, their value is very close to each other.
- Recall that $V(\hat{\beta}_1) = \frac{\sigma^2}{n} E\left[\frac{1}{S_X^2}\right]$
  - We estimate $V(\hat{\beta}_1)$ by $\hat{\sigma}^2 \Rightarrow \hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{nS_X^2}$, where $E\left[\frac{1}{S_X^2}\right]$ is estimated by $\frac{1}{S_X^2}$
- Recall $V(\hat{\beta}_0) = \sum_i \frac{X_i^2}{n} V(\hat{\beta}_1)$
  - We estimate $V(\hat{\beta}_0)$ by $\hat{\sigma}^2 \Rightarrow \hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum_i X_i^2}{n^2 S_X^2}$

The Goodness of Fit Measures

- The goodness of fit measures are the measures of how well our model fits the data.
- **Standard error of regression**, $\hat{\sigma}$, is one intuitive way of approaching goodness of fit.
- Recall that in the standard simple linear regression, we have $Y = \beta_0 + \beta_1 X + \varepsilon$ and $V(\varepsilon|X) = \sigma^2$.
- One problem with this goodness of fit method is that its magnitude heavily depends on the units of Y

The Goodness of Fit Measures

- **Coefficient of Determination**, $\mathbf{R^2}$, is a very common measure of goodness of fit

$$R^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i \varepsilon_i^2}{\sum_i y_i^2} \tag{1}$$

  - where $y_i = Y_i - \bar{Y}$, $\hat{y}_i = \hat{Y}_i - \bar{Y}$, and $\varepsilon_i = Y_i - \hat{Y}_i$
  - Notice that $y_i = \hat{y}_i + \varepsilon_i$. This implies that $0 \leq \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} \leq 1$
  - If $R^2 = 1$, the model explains all of the variation (100%) in Y
  - If $R^2 = 0$, the model explains none of the variation (0%) in Y

The Goodness of Fit Measures

- A low R^2 implies:
  - The model fails to explain a significant proportion of the variation of Y.
  - A low R2 does **not** necessarily mean that the estimates are unhelpful or inappropriate.

- The R^2 can be used for comparing different models for the same dependent variable Y.

- Also, for the correlation coefficient between actual and estimated values of Y, $\rho_{Y,\hat{Y}}$, we ahe the following relationship: $R^2 = \left(\hat{\rho}_{Y,\hat{Y}}\right)^2$