

Lecture Note 1: Two Variable Regression Model

Economics 143

Jinyong Hahn

Comments by Michael Powell

Fitting a Line

- Data:

Y (GPA)	X (family income in \$1,000)
4.0	21.0
3.0	15.0
3.5	15.0
2.0	9.0
3.0	12.0
3.5	18.0
2.5	6.0
2.5	12.0

- Given a scatter diagram of (X, Y) , we want to find the best linear relation between X and Y .
- Least Squares Criterion:

$$\min_{a,b} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

where

$$\hat{Y}_i = a + bX_i$$

Here, N denotes the number of observation, and \hat{Y}_i denotes the fitted value.

- In words, the least squares criterion minimizes the sum of squared deviations of the fitted values from actual value of Y .
- Least squares solution:

$$b = \frac{N \sum_{i=1}^N X_i Y_i - \sum_{i=1}^N X_i \sum_{i=1}^N Y_i}{N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2}, \quad a = \bar{Y} - b\bar{X}$$

Proof

$$\begin{aligned} & \min_{a,b} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \\ &= \min_{a,b} \sum_{i=1}^N (Y_i - a - bX_i)^2 \end{aligned}$$

Taking first order conditions, we have:

$$(a) : 2 \sum_{i=1}^N (Y_i - a - bX_i) (-1) = 0 \quad (1)$$

$$(b) : 2 \sum_{i=1}^N (Y_i - a - bX_i) (-X_i) = 0 \quad (2)$$

Rewriting (1), we have:

$$\begin{aligned} \sum_{i=1}^N (Y_i - a - bX_i) &= 0 \\ \sum_{i=1}^N Y_i - Na - b \sum_{i=1}^N X_i &= 0 \\ Na &= \sum_{i=1}^N Y_i - b \sum_{i=1}^N X_i \\ a &= \frac{1}{N} \sum_{i=1}^N Y_i - \frac{b}{N} \sum_{i=1}^N X_i \\ a &= \bar{Y} - b\bar{X} \end{aligned}$$

Rewriting (2), we have:

$$\begin{aligned} \sum_{i=1}^N (Y_i X_i - aX_i - bX_i^2) &= 0 \\ \sum_{i=1}^N Y_i X_i - a \sum_{i=1}^N X_i - b \sum_{i=1}^N X_i^2 &= 0 \\ b \sum_{i=1}^N X_i^2 &= \sum_{i=1}^N Y_i X_i - a \sum_{i=1}^N X_i \end{aligned}$$

If we plug in $a = \bar{Y} - b\bar{X}$, we get:

$$\begin{aligned}
b \sum_{i=1}^N X_i^2 &= \sum_{i=1}^N Y_i X_i - (\bar{Y} - b\bar{X}) \sum_{i=1}^N X_i \\
&= \sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i + b\bar{X} \sum_{i=1}^N X_i \\
b \left(\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i \right) &= \sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i \\
b &= \frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i} \\
&= \frac{\sum_{i=1}^N Y_i X_i - \left(\frac{1}{N} \sum_{i=1}^N Y_i \right) \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \left(\frac{1}{N} \sum_{i=1}^N X_i \right) \sum_{i=1}^N X_i} \\
&= \frac{N \sum_{i=1}^N Y_i X_i - \sum_{i=1}^N Y_i \sum_{i=1}^N X_i}{N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2}
\end{aligned}$$

- Least squares line passes through sample means of Y and X :

$$\bar{Y} = a + b\bar{X}$$

Proof Recall that $Y_i = a + bX_i$, where

$$b = \frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i}$$

and

$$a = \bar{Y} - \bar{X} \left(\frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i} \right)$$

Therefore, we have:

$$Y_i = \bar{Y} - \left(\frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i} \right) \bar{X} + \left(\frac{\sum_{i=1}^N Y_i X_i - \bar{Y} \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i^2 - \bar{X} \sum_{i=1}^N X_i} \right) X_i$$

Evaluating at $X_i = \bar{X}$, we have:

$$\begin{aligned}
Y_i &= \bar{Y} - \left(\frac{\sum_{i=1}^N Y_i \bar{X} - \bar{Y} \sum_{i=1}^N \bar{X}}{\sum_{i=1}^N \bar{X}^2 - \bar{X} \sum_{i=1}^N \bar{X}} \right) \bar{X} + \left(\frac{\sum_{i=1}^N Y_i \bar{X} - \bar{Y} \sum_{i=1}^N \bar{X}}{\sum_{i=1}^N \bar{X}^2 - \bar{X} \sum_{i=1}^N \bar{X}} \right) \bar{X} \\
&= \bar{Y}
\end{aligned}$$

That is, the OLS estimator passes through the point (\bar{X}, \bar{Y}) .

- Alternative representation:

$$b = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}$$

where

$$\begin{aligned} x_i &= X_i - \bar{X} \\ y_i &= Y_i - \bar{Y} \end{aligned}$$

Proof Here, my goal is to show that this alternative representation is, indeed, equivalent to the standard representation.

$$\begin{aligned} b &= \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} \\ &= \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^N (X_i Y_i - Y_i \bar{X} - X_i \bar{Y} + \bar{X} \bar{Y})}{\sum_{i=1}^N (X_i^2 - 2X_i \bar{X} + \bar{X}^2)} \\ &= \frac{\sum_{i=1}^N X_i Y_i - \bar{X} \sum_{i=1}^N Y_i - \bar{Y} \sum_{i=1}^N X_i + N \bar{X} \bar{Y}}{\sum_{i=1}^N X_i^2 - 2\bar{X} \sum_{i=1}^N X_i + N \bar{X}^2} \end{aligned}$$

Recall that $\sum_{i=1}^N X_i = N\bar{X}$ and $\sum_{i=1}^N Y_i = N\bar{Y}$. This gives us:

$$\begin{aligned} b &= \frac{\sum_{i=1}^N X_i Y_i - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y}}{\sum_{i=1}^N X_i^2 - 2N\bar{X}^2 + N\bar{X}^2} \\ &= \frac{\sum_{i=1}^N X_i Y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^N X_i^2 - N\bar{X}^2} \\ &= \frac{\sum_{i=1}^N X_i Y_i - N \left(\frac{1}{N} \sum_{i=1}^N X_i \right) \left(\frac{1}{N} \sum_{i=1}^N Y_i \right)}{\sum_{i=1}^N X_i^2 - N \left(\frac{1}{N} \sum_{i=1}^N X_i \right)^2} \\ &= \frac{N \sum_{i=1}^N X_i Y_i - \left(\sum_{i=1}^N X_i \right) \left(\sum_{i=1}^N Y_i \right)}{N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2} \end{aligned}$$

Model

- For a given value of X (independent variable), we may observe many possible values of Y (dependent variable)
- Example: Consumption of an individual who earns \$50,000 a year?
- Mathematical model:

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$$

where Y is a random dependent variable, X is a fixed (nonstochastic) independent variable, and ε is a random error term

- Why error?: Model is a simplification of reality. There could be omitted variables related to consumption, e.g., individual tastes. Or, there could be a measurement error

Assumptions

1. Linear relation:

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$$

2. X are nonstochastic
3. For all i , $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2$
4. ε_i are independent: $E[\varepsilon_i \varepsilon_j] = 0$
5. ε_i is normally distributed

Remark

- Assumptions 1 - 4: Classical Linear Regression Model
- X are nonstochastic = X are under control by researcher = Experiment!
- $E[\varepsilon_i] = 0$: Matter of convenience. If $E[\varepsilon_i] = \alpha'$, we can rewrite

$$\begin{aligned} Y_i &= \alpha + \beta \cdot X_i + \varepsilon_i - \alpha' \\ &= \alpha^* + \beta \cdot X_i + \varepsilon_i^* \end{aligned}$$

- $Var[\varepsilon_i]$ is constant: Homoscedasticity.
- Violation of homoscedasticity is called heteroscedasticity
- Equivalent specification:

$$\begin{aligned} E[Y_i] &= \alpha + \beta \cdot X_i \\ Var[Y_i] &= \sigma^2 \\ &Y_i \text{ are independent} \\ Y_i &\sim N(\alpha + \beta \cdot X_i, \sigma^2) \end{aligned}$$

Least Squares Estimation

- Recall least squares formula:

$$b = \frac{\sum x_i y_i}{\sum x_i^2}, \quad a = \bar{Y} - b\bar{X}$$
$$x_i = X_i - \bar{X}, \quad y_i = Y_i - \bar{Y}$$

- *Gauss-Markov Theorem*: Given Assumptions 1-4, the least squares estimators a and b are the best (most efficient) linear unbiased estimators of α and β : a and b have the minimum variance of all linear unbiased estimators
- A linear estimator is an estimator which can be written as $\sum c_i Y_i$ for some fixed numbers c_i
- b is a linear estimator: It can be written as a weighted average of individual observations on Y

$$b = \sum w_i Y_i$$

where

$$w_j = \frac{x_j}{\sum x_i^2}$$

Proof

$$\begin{aligned} \frac{\sum_{i=1}^n x_i y_i}{\sum_{j=1}^n x_j^2} &= \frac{\sum_{i=1}^n x_i (Y_i - \mu_Y)}{\sum_{j=1}^n x_j^2} \\ &= \frac{\sum_{i=1}^n x_i Y_i}{\sum_{j=1}^n x_j^2} - \frac{\sum_{i=1}^n x_i \mu_Y}{\sum_{j=1}^n x_j^2} \\ &= \sum_{i=1}^n w_i Y_i - \mu_Y \sum_{i=1}^n w_i \\ &= \sum_{i=1}^n w_i Y_i \end{aligned}$$

Where in the last step, I recognized that $\sum_{i=1}^n w_i = 0$. The proof of this property is below.

- a and b have the smallest possible variances among all linear unbiased estimators
- Least squares estimators are BLUEs (Best Linear Unbiased Estimator)
- Gauss-Markov Theorem does not apply to nonlinear estimators: There may exist a nonlinear estimator with a smaller mean squared error
- Some useful facts:

Proposition $\sum w_i = 0$

Proof

$$\begin{aligned}\sum_{j=1}^N w_j &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} \\ &= \frac{\sum_{j=1}^N x_j}{\sum_{i=1}^N x_i^2} \\ &= \frac{\sum_{j=1}^N (X_j - \bar{X})}{\sum_{i=1}^N x_i^2} \\ &= \frac{\sum_{j=1}^N X_j - \sum_{j=1}^N \bar{X}}{\sum_{i=1}^N x_i^2} \\ &= \frac{N\bar{X} - N\bar{X}}{\sum_{i=1}^N x_i^2} = 0\end{aligned}$$

Proposition $\sum w_i X_i = \sum w_i x_i = 1$

Proof

$$\begin{aligned}\sum_{j=1}^N w_j x_j &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} x_j \\ &= \frac{\sum_{j=1}^N x_j^2}{\sum_{i=1}^N x_i^2} \\ &= \frac{\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i^2} = 1\end{aligned}$$

$$\begin{aligned}\sum_{j=1}^N w_j x_j &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} (X_j - \bar{X}) \\ &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} X_j - \bar{X} \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} \\ &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} X_j - \bar{X} \sum_{j=1}^N w_j \\ &= \sum_{j=1}^N \frac{x_j}{\sum_{i=1}^N x_i^2} X_j = \sum_{j=1}^N w_j X_j\end{aligned}$$

Proposition $\sum w_i^2 = \frac{1}{\sum x_i^2}$

Proof

$$\begin{aligned}\sum_{j=1}^N w_j^2 &= \sum_{j=1}^N \left(\frac{x_j}{\sum_{i=1}^N x_i^2} \right)^2 \\ &= \sum_{j=1}^N \frac{x_j^2}{\left(\sum_{i=1}^N x_i^2 \right)^2} \\ &= \frac{\sum_{j=1}^N x_j^2}{\left(\sum_{i=1}^N x_i^2 \right)^2} \\ &= \frac{1}{\sum_{i=1}^N x_i^2}\end{aligned}$$

Proposition $b = \beta + \sum w_i \varepsilon_i$

Proof

$$\begin{aligned}b &= \sum_{i=1}^N w_i Y_i \\ &= \sum_{i=1}^N w_i (\alpha + \beta X_i + \varepsilon_i) \\ &= \underbrace{\alpha \sum_{i=1}^N w_i}_{=0} + \beta \underbrace{\sum_{i=1}^N w_i X_i}_{=1} + \sum_{i=1}^N w_i \varepsilon_i \\ &= \beta + \sum_{i=1}^N w_i \varepsilon_i\end{aligned}$$

Proposition $a = \alpha + \sum \left(\frac{1}{N} - \bar{X} w_i \right) \varepsilon_i$

Proof

$$\begin{aligned}a &= \bar{Y} - b \bar{X} \\ &= \bar{Y} - \bar{X} \left(\beta + \sum_{i=1}^N w_i \varepsilon_i \right) \\ &= \bar{Y} - \beta \bar{X} - \bar{X} \sum_{i=1}^N w_i \varepsilon_i\end{aligned}$$

Recall that $\bar{Y} = \alpha + \beta \bar{X}$. This gives us:

$$\begin{aligned}a &= \alpha + \beta \bar{X} - \beta \bar{X} - \bar{X} \sum_{i=1}^N w_i \varepsilon_i \\ &= \alpha - \bar{X} \sum_{i=1}^N w_i \varepsilon_i\end{aligned}$$

$$\begin{aligned}
a &= \alpha + \sum_{i=1}^N \left(\frac{1}{N} - \bar{X}w_i \right) \varepsilon_i \\
&= \alpha + \sum_{i=1}^N \left(\frac{\varepsilon_i}{N} - \bar{X}w_i\varepsilon_i \right) \\
&= \alpha + \sum_{i=1}^N \frac{\varepsilon_i}{N} - \bar{X} \sum_{i=1}^N w_i\varepsilon_i \\
&= \alpha + \frac{1}{N} \sum_{i=1}^N \varepsilon_i - \bar{X} \sum_{i=1}^N w_i\varepsilon_i
\end{aligned}$$

Properties of Least Squares Estimators

- b is normally distributed: It is a linear combination of normally distributed random variables.

$$b = \sum w_i Y_i = \beta + \sum w_i \varepsilon_i$$

- All we need to characterize its distribution are mean and variance:

$$\begin{aligned}
E[b] &= E \left[\beta + \sum w_i \varepsilon_i \right] \\
&= \beta + \sum w_i E[\varepsilon_i] \\
&= \beta
\end{aligned}$$

$$\begin{aligned}
Var[b] &= Var \left[\beta + \sum w_i \varepsilon_i \right] \\
&= \sum w_i^2 \cdot Var[\varepsilon_i] \\
&= \sigma^2 \sum w_i^2 \\
&= \sigma^2 / \sum x_i^2
\end{aligned}$$

- To summarize, we have

$$b \sim N \left(\beta, \frac{\sigma^2}{\sum x_i^2} \right)$$

- With similar argument, we can show that

$$a \sim N \left(\alpha, \sigma^2 \frac{\sum X_i^2}{N \sum x_i^2} \right)$$

- We can also show that

$$Cov(a, b) = -\frac{\bar{X}\sigma^2}{\sum x_i^2}$$

- Observe that β is the marginal response of Y with respect to a unit change in X . What is the accuracy of its estimator?

$$\sigma^2 / \sum x_i^2$$

Note that it is small if σ^2 is small and $\sum x_i^2$ is large.

Estimation of Variances

- We usually do not know σ^2
- It can be shown that

$$s^2 = \frac{1}{N-2} \sum e_i^2 = \frac{1}{N-2} \sum (Y_i - a - b \cdot X_i)^2$$

is an unbiased estimator of σ^2

- Here,

$$e_i = Y_i - a - b \cdot X_i$$

is sometimes called the residual

- Note that the denominator or the degree of freedom is now $N - 2$.
- Estimators of the variances of a and b are constructed by replacing σ^2 with s^2 :

$$\begin{aligned} s_b^2 &= \frac{s^2}{\sum x_i^2} \\ s_a^2 &= s^2 \frac{\sum X_i^2}{N \sum x_i^2} \\ \widehat{Cov}(a, b) &= -\frac{\bar{X} s^2}{\sum x_i^2} \end{aligned}$$

Inference

- Confidence interval: It is known that

$$\frac{b - \beta}{s_b} \sim t(N - 2)$$

Therefore, a valid confidence interval for β is given by

$$b \pm t_c(N - 2) \cdot s_b$$

Similarly, a valid confidence interval for β is given by

$$a \pm t_c(N - 2) \cdot s_a$$

- Example 3.1:

$$b = .12, \quad N = 8, \quad \sum_{i=1}^N e_i^2 = .6528, \quad s^2 = \frac{.6528}{8-2} = 1.09, \quad \sum_{i=1}^N x_i^2 = 162,$$

$$s_b^2 = \frac{1.09}{162}, \quad s_b = \sqrt{\frac{1.09}{162}} = .0259$$

It can be seen that $t_c(6) = 2.447$. Therefore, 95% confidence interval is given by

$$.12 \pm 2.447(.0259) = .12 \pm .06$$

- Hypothesis testing of $H_0 : \beta = \beta_0$: Reject the null if

$$\left| \frac{b - \beta_0}{s_b} \right| > t_c(N - 2)$$

Equivalently, reject the null if β_0 is not contained in confidence interval. Similar strategy for α

- Example 3.1 with $H_0 : \beta = 0$: t-statistic equals

$$\left| \frac{.12 - 0}{.0259} \right| = 4.6 > 2.447$$

We reject the null. Equivalently, we observe that confidence interval $.12 \pm .06$ does not contain 0 and reject the null.

R^2

- Define

$$\text{variation}(Y) = \sum (Y_i - \bar{Y})^2$$

- Observe that

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

- It can be shown that

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \\ TSS &= ESS + RSS \end{aligned}$$

- TSS: Total Sum of Squares
- ESS: Error Sum of Squares
- RSS: Regression Sum of Squares

- Observe that

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

- We define

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

- R^2 is the proportion of the total variation of Y explained by regression on X
- It can be shown that

$$0 \leq R^2 \leq 1$$

and

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

- R^2 is a descriptive statistic. Roughly speaking, we associate a high value of R^2 with a good fit of regression, and a low value of R^2 with a poor fit. No formal statistical interpretation is attached, though.

Capital Asset Pricing Model: To be used in Problem Set

- Linear relationship between risk and return
- Rate of return r of an investment is defined by

$$r = \frac{p_1 + d - p_0}{p_0}$$

where p_1 , d , and p_0 denote the price of the security at the end of the time period, the dividend paid during the time period, and the price of the security at the beginning of the time period, respectively.

- Return is easily calculated ex post (once the investment has been made), but is uncertain ex ante (before the investment decision has been made).
- If investors were to purchase an asset having zero risk, they would still demand a return as an inducement to postpone current consumption. Such a return is called the risk-free rate of return, and we denote it by r_f .
- Risk premium of the j th asset is defined to be

$$r_j - r_f$$

- Given a portfolio a with expected return r_a and variance σ_a^2 , a new portfolio constructed by combining a with a risk free asset has expected return equal to

$$r_p = (1 - w_a) \cdot r_f + w_a \cdot r_a$$

and variance equal to

$$\sigma_p^2 = w_a^2 \cdot \sigma_a^2,$$

where w_a denotes the porportion of total funds invested in portfolio a . Therefore, we have

$$r_p - r_f = w_a \cdot (r_a - r_f) = \frac{\sigma_p}{\sigma_a} \cdot (r_a - r_f).$$

- Now, consider a small portfolio whose sole security is asset j . It can be shown via some lengthy argument that

$$r_j - r_f = \frac{\sigma_{jm}}{\sigma_m^2} \cdot (r_m - r_f),$$

where $r_m, \sigma_m^2, \sigma_{jm}$ denote the return of the entire market portfolio, the variance of market portfolio, and the covariance between j and market portfolio, respectively.

- This is the celebrated CAPM model, which can be rewritten as a simple linear regression model:

$$r_j - r_f = \alpha_j + \beta_j \cdot (r_m - r_f) + \varepsilon_j$$