Simple Linear Regression Model

ECON 340: Economic Research Methods — Instructor: Div Bhagia

We are interested in the relationship between two variables, X and Y. Here, Y is the *dependent variable*, while X is the *independent or explanatory variable*.

1 Ordinary Least Squares (OLS)

We observe Y_i and X_i for all individuals in our sample. We want to fit a line to represent our data as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

where $\hat{Y}_i$ is the predicted value of the dependent variable. $\hat{\beta}_0$ is the intercept for this line, while $\hat{\beta}_1$ is its slope. $\hat{\beta}_0$ and $\hat{\beta}_1$ are also called regression coefficients. The OLS estimator chooses the regression coefficients to minimize the discrepancy between $\hat{Y}_i$ and Y_i. Define residuals as

$$\hat{u}_i = Y_i - \hat{Y}_i$$

Then the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by minimizing the sum of squared residuals:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} \hat{u}_i^2 = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Some calculus (included at the end of this handout) will reveal

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

Note that from the expression of $\hat{\beta}_0$, the best fit line passes through the point of means.

2 Goodness of Fit: The R^2

Once we have estimated a regression line, we might be interested in how well this line fits the data. The R-squared measures how well the OLS regression line fits the data. R-squared is the percent of sample variation in Y that is explained by X. Note that,

$$Y_i = \hat{Y}_i + \hat{u}_i$$

In this notation, R^2 is the ratio of sample variation of $\hat{Y}_i$ to sample variation of Y_i. Before we write down the formula for R^2, let's introduce a few more terms.

Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Explained Sum of Squares:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

I am not going to prove it, but $TSS = ESS + RSS$. RSS is also sometimes called the *Sum of squared residuals* (SSR).

A measure of goodness of fit:

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

R^2 lies between 0 and 1

- If X explains no variation in Y, $\hat{\beta}_1 = 0$ and $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$. In which case, $ESS = 0$ and hence $R^2 = 0$.
- On the other hand, if X explains all the variation in Y, $\hat{Y}_i = Y_i$ and $RSS = 0$. In

which case, $R^2 = 1$.

3 Linear Regression Model: Assumptions for Causal Inference

Until now we have talked about choosing a line that fits the sample data without any reference to the underlying population. We will now formally set up the linear regression model and discuss the assumptions under which OLS estimates can be used to answer causal questions.

- *Assumption 1 (Linear in Parameters)*: The population regression model is linear in its parameters and correctly specified as:

$$Y = \beta_0 + \beta_1 X + u$$

  Here, u is the mean zero error term $E(u) = 0$. Note that the model can be non-linear in variables. For example, $Y = \beta_0 + \beta_1 X^2 + u$ or $\ln Y = \beta_0 + \beta_1 \ln X + u$ are fine.

- *Assumption 2 (Random Sample):* The observed data (Y_i, X_i) for $i = 1, 2, ..., n$ represent a random sample of size n from the above population model.

- *Assumption 3 (No large outliers)*: Fourth moments (or Kurtosis) of X and Y are finite.

- *Assumption 4 (Zero Conditional Mean/Exogeneity):* The expected value of the error term is 0 conditional on any value of the explanatory variable.

$$E(u|X) = 0$$

Assumptions 1-4 imply that OLS estimators are unbiased, that is

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$$

Note that Assumption 4 is the key assumption that is needed for causal analysis and the one most often not satisfied in practice. By Assumption 4,

$$E(Y|X) = \beta_0 + \beta_1 X$$

Say $X = x$ and it increases by 1 unit. Note that,

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

$$E(Y|X = x + 1) = \beta_0 + \beta_1(x + 1)$$

Then β_1 represents the causal effect of one unit change in X on Y

$$\beta_1 = E(Y|X = x + 1) - E(Y|X = x)$$

Note that the error term u captures unobserved factors that may affect the outcome Y. The exogeneity assumption posits that, on average, these unobserved factors do not vary with the values of X. In other words, omitted factors are uncorrelated with X.

4 Sampling Distribution of OLS Estimators

Since OLS estimators are computed from a random sample, just like the sample mean, they are random variables. In small samples, the sampling distribution for OLS estimators is a bit complicated, but in large samples, they are approximately normal because of the Central Limit Theorem (CLT).[1]

Under assumptions 1-4, in large samples,

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}), \qquad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

where

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{Var[(X_i - \mu_X)u_i]}{Var(X_i)}$$

[1] In small samples, if we are willing to assume that errors are conditionally normally distributed, we can conclude that OLS estimators are normally distributed.

5 Hypothesis Testing and Confidence Intervals

As before, we don't have $\sigma^2_{\hat{\beta}_1}$ but we can estimate its sample counterpart $S_{\hat{\beta}_1}$. Then we can compute the t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

Since $\hat{\beta}_1$ is approximately normally distributed in large samples, the t-statistic is approximately distributed as a standard normal random variable.

One common hypothesis test we are interested in involves testing whether the effect is 0. This is so common that most statistical packages automatically report the t and p values associated with this hypothesis test.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

In this case the test statistic is

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

As before, denote $z_{\alpha/2}$ as the value of z that leaves area $\alpha/2$ in the upper tail of the normal distribution. If $|t| > z_{\alpha/2}$, we can reject the null at α% level of significance and say that β is statistically significant at α% level of significance. Alternatively, if our p-value associated with this test is smaller than α, we can say that β is statistically significant at α% level of significance.

Similarly, we can construct a $1 - \alpha$% confidence interval (CI) around the β. A $1 - \alpha$ CI in this case is given by:

$$\hat{\beta}_1 \pm z_{\alpha/2} \cdot S_{\hat{\beta}_1}$$

where $z_{\alpha/2}$ leaves area $\alpha/2$ in the upper tail of the normal distribution.

Appendix: Derivation of OLS Coefficients[2]

Note that,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

First-order conditions for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$-2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{1}$$

$$-2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \tag{2}$$

Dividing both sides of equation (1) by $-2n$, we get

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

This implies that the best fit line will pass through the point of means. We can rewrite the above equation to get

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

Now if we both sides of equation (2) by -2 and plug in $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, we get

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)X_i = 0 \\
&\Rightarrow \sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})]X_i = 0 \\
&\Rightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})X_i = \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})X_i \\
&\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})X_i} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}
\end{aligned}$$

[2] You do not need to know the derivation for the purpose of the exam. However, you do need to know how $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen.